Montana Comprehensive Assessment System
(MontCAS, Phase 2)
Criterion-Referenced Test Alternate Assessment
(CRT-Alternate)

CRT-ALTERNATE
TECHNICAL MANUAL
2004

ITEM ANALYSIS

As noted in Brown (1983), “a test is only as good as the items it contains.” A complete evaluation of a test’s quality must include an evaluation of each question. Both the *Standards for Educational and Psychological Testing* and the *Code of Fair Testing Practices in Education* include standards for identifying quality questions. Questions should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, questions must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that Montana CRT-Alternate questions meet these standards. This report focuses on the more quantitative evaluations. The statistical evaluations included are: difficulty indices, item-test correlations, and differential item functioning (DIF) analyses. The item analyses presented here are based on the statewide administration of the Montana CRT-Alternate in spring 2004. About 80 grade 4 students, 90 grade 8 students, and 100 grade 10 students participated in the assessment.

Difficulty Indices (*p*)

All tasks were evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty was defined as the average proportion of points achieved on an item, and was measured by obtaining the average score on an item and dividing by the maximum score for the item. Tasks are scored polytomously, where a student can achieve a score of 0, 1, 2, 3, or 4 for the item. By computing the difficulty index as the average proportion of points achieved, the items are placed on a scale that ranges from 0.0 to 1.0. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an “easiness index” because larger values indicate easier questions.

An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item. Ideally, the items on an assessment will have a range of difficulties between 0.25 and 0.9 with most items falling between 0.4 and 0.7. Items that have either a very high or very low difficulty index are considered to be potentially problematic because they are either so difficult that few students get them right or so easy that nearly all students get them right. In either case, such items should be reviewed for appropriateness for inclusion on the assessment. If an assessment were comprised entirely of very easy or very hard items, all students would receive nearly the same scores and the assessment would not be able to differentiate high ability students from low ability students.

Item-Test Correlations (Item Discrimination)

A desirable feature of an item is that the higher ability students perform better on the item than lower ability students. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of an item. Within classical test theory, the item-test correlation is referred to as the item's discrimination because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. The discrimination index used to evaluate Montana CRT-Alternate tasks was the Pearson product-moment correlation. The theoretical range of these statistics is –1 to +1, with a typical range from .3 to .6.

Discrimination indices can be thought of as measures of how closely a question assesses the same knowledge and skills assessed by other questions contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. For the Montana CRT-Alternate, the test total score was used as the criterion score.

Summary of Item Analysis Results

A summary of the item difficulty and item discrimination statistics for each grade/content combination is presented in Table 1 below.

Table 1
Item Analysis

Grade	Content Area	Difficulty		Discrimination	
		Mean	StDev	Mean	StDev
4	Reading	0.75	0.17	0.67	0.20
	Mathematics	0.70	0.16	0.75	0.17
8	Reading	0.78	0.12	0.77	0.11
	Mathematics	0.63	0.17	0.80	0.14
10	Reading	0.79	0.10	0.75	0.16
	Mathematics	0.71	0.13	0.79	0.17

Differential Item Functioning

Investigations of item or test bias seek to determine whether scores for subgroups of students may be affected by attributes other than those the test is intended to measure. Such investigations usually begin by examining whether subgroups of students perform differently than expected on individual items. Specifically, differences due to irrelevant factors are examined. If such differential item functioning (DIF) is detected, a qualitative assessment of the item is made to determine whether the item is biased against a particular group. It should be noted that the detection of DIF does not imply that the item is biased; instead, it is a statistical tool that helps identify items that may be biased.

Investigations of test fairness, in contrast to bias, seek to determine whether the test predicts academic success equally well for minorities and non-minorities. Although these concepts are related, the first is generally considered a measurement issue, while the second is a legal issue.

The *Code of Fair Testing Practices in Education* explicitly states that subgroup differences in performance due to irrelevant factors should be examined when sample size permits, and actions should be taken to make certain that differences in performance are due to construct-relevant, rather than irrelevant, factors. The *Standards for Educational and Psychological Testing* includes similar guidelines.

DIF procedures are designed to identify questions for which subgroups of interest perform differentially beyond the impact of differences in overall achievement. However, due to very small sample sizes (i.e., around 100 total students) it is unreasonable to calculate DIF statistics for the Montana CRT-Alternate. That is, Type I error rates would be unreasonably high and would result in incorrect conclusions regarding the functioning of the items between reference and focal groups. Thus, DIF statistics are not included as part of this technical report.

RELIABILITY

Although an individual question's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way questions function together and complement one another. Tests that function well provide an accurate assessment of the student's level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either higher or lower than his or her true ability. For example, a student may mis-read a question, or mistakenly bubble in the wrong bubble when he or she knew the answer; similarly a student may get a question correct by guessing, even though he or she did not know the answer. Collectively, these extraneous factors that impact a student's score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement can be perfectly accurate. This is true of academic assessments—no assessment can measure students perfectly accurately; some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. When tests have a high amount of measurement error student scores are very unstable. Students with high ability may get low scores or vice versa. Consequently, one cannot reliably tell a student's true level of ability with such a test. Questions that function well together produce assessments that have less measurement error; that is, the errors made should be small on average and student scores on such a test will consistently represent their ability. Such assessments are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable (this is referred to as test-retest reliability). A potential problem with this approach is that students may remember questions from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the 'remembering questions' problem is to give a different, but parallel test at the second administration. If student scores on each test correlate highly the test is considered reliable (this is known as alternate forms reliability, because an alternate form of the test is used in each administration). This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. One way to address these problems is to split the test in half and then correlate students' scores on the two half-tests; this in effect treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval are alleviated. This is known as a split-half estimate of reliability. If the two half-test scores correlate highly, questions on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the questions complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires a judgement regarding the selection of which questions contribute to which half-test score. This decision may have an impact on the resulting correlation; different splits will give different estimates of reliability. Cronbach (1951) provided a statistic, α, that avoids this concern about the split-half method. Cronbach's α gives an estimate of the average of all possible splits for a given test.

Cronbach's α is often referred to as a measure of internal consistency because it provides a measure of how well all the items in the test measure one single underlying ability.

Reliability and Standard Errors of Measurement

Table 2 presents Cronbach's α coefficient for each subject area (reading and mathematics), for each grade level.

Table 2
Reliability Analysis – All Grades

Grade	Content Area	Reliability
4	Mathematics	0.98
	Reading	0.96
8	Mathematics	0.98
	Reading	0.97
10	Mathematics	0.98
	Reading	0.97

SCALING

Translating Raw Scores to Scaled Scores and Performance Levels

Montana CRT-Alternate scores in each content area are reported on a scale that ranges from 200 to 300. Scaled scores supplement the Montana CRT-Alternate performance-level results by providing information about the position of a student's results within a performance level. School- and district-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores, or total number of points, on the Montana CRT-Alternate tests are translated to scaled scores using a data analysis process called **scaling**. Scaling simply converts raw points from one scale to another. In the same way that the same temperature can be expressed on either the Fahrenheit or Celsius scales and the same distance can be expressed either in miles or kilometers, student scores on the Montana CRT-Alternate tests could be expressed as raw scores (i.e., number right) or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change the students' performance-level classifications. Given the relative simplicity of raw scores, it is fair to question why scaled scores are used in Montana CRT-Alternate reports instead of raw scores. Foremost, scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Because the standard-setting process typically results in different cut scores across content areas on a raw score basis, it is useful to transform these raw cut scores to a scale that is more easily interpretable and consistent. For the Montana CRT-Alternate, a score of 225 is the cut score between the **Novice** and **Nearing Proficiency** performance levels. This is true regardless of which content area, grade, or year one may be concerned with. If one were to use raw scores, the raw cut score between **Novice** and **Nearing Proficiency** may be, for example, 35 in mathematics at grade 8, but may be 33 in mathematics at grade 11, or 36 in reading at grade 8. Using scaled scores greatly simplifies the task of understanding how a student performed.

As previously stated, student scores on the Montana CRT-Alternate are reported in integer values from 200 to 300 with three scores representing cut scores on each assessment. Table 3 presents the scaled score range for each performance level in each grade-content area combination. The determination of these cut scores is detailed in the Montana CRT-Alternate standard setting report.

Table 3

Grade	Content Area	Scaled Score Range for each Performance Level			
		Novice	Nearing proficiency	Proficient	Advanced
4	Reading	200–224	225–249	250–267	268–300
	Mathematics	200–224	225–249	250–277	278–300
8	Reading	200–224	225–249	250–262	263–300
	Mathematics	200–224	225–249	250–268	269–300
10	Reading	200–224	225–249	250–266	267–300
	Mathematics	200–224	225–249	250–275	276–300

The scaled scores are obtained by a simple linear transformation of the □s (for Montana CRT-Alternate, the raw scores) using the values of 225 and 250 on the scaled score metric and the □ values obtained through standard setting to define the transformation. For example, the following equation was used to determine the scaled scores for each.

$$ss = m(\Theta) + b$$

where

$$m = (225 - 250)/(\Theta_1 - \Theta_2),$$

$$b = 225 - m(\Theta_1)$$

and SS is the scaled score value, $□_1$ is the cut score on the □ metric for the novice/nearing proficiency cut and $□_2$ is the cut score on the □ metric for the nearing proficiency/proficient cut. In this equation, *m* represents the slope of the line providing the relationship between Θ and the scaled scores. The scaled score values of 225 and 250 were used because they are the scaled score cut points between novice and nearing proficiency and nearing proficiency and proficient, respectively. The determination of $□_1$ and $□_2$ is detailed in the Montana CRT-Alternate standard setting report.

REFERENCES

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth, TX: Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Joint Committee on Testing Practices (1988). *Code of fair testing practices in education*. Washington, DC: National Council on Measurement in Education.